DOCUMENT RESUME

ED 279 669 TM 870 056

AUTHOR Coffman, William E.
TITLE Recommendations on Writing Assessments for Future NAEPs.
PUB DATE 22 Sep 86
NOTE 7p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Age Differences; *Educational Assessment; Educational Objectives; Educational Testing; Elementary Secondary Education; Holistic Evaluation; *National Surveys; Scaling; Scoring; *Test Construction; *Testing Problems; Testing Programs; *Test Results; *Writing Evaluation
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

These comments provide further details which expand upon the recommendations of the subcommittee on National Assessment of Educational Progress (NAEP) Writing. Three issues are discussed. The first concerns how objectives of assessment in writing ought to be stated. The 1984 objectives are stated in general terms without reference to what might reasonably be expected of individuals in a particular age group or at different levels of ability within age groups. They differ from earlier objectives by emphasizing process rather than product, a focus which does not provide the type of information desired by NAEP's audience. The second issue is the criteria for the selection of exercises for several different age and grade levels. The content of some of the items and some of the conditions under which tests were administered were not realistic. For example, writers, in real-life situations, are usually allowed access to resource materials and time to revise their work. The third area of concern is the reporting of results. Here, five recommendations are made: (1) replace the current Primary Trait and Global Rating scales with anchored rating scales; (2) provide more information relating responses to objectives; (3) examine the response patterns for pairs of items; (4) abandon the ARM scaling; and (5) examine the effects of dependencies across grouping categories on the magnitude of group differences. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279669

Recommendations on Writing Assessments for Future NAEPs

William E. Coffman

University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper Commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

BEST COPY AVAILABLE

TM 870 058

M E M O R A N D U M

FOR: H. Thomas James
cc: Eva Baker, Edys Quellmals, Paul Williams
From William C. Coffman, 201 West First Ave Apt #6L Iowa City, IA 52240
SUBJECT: Recommendations on Writing Assessments for Future NAEPs

DATE: September 22, 1986

Eva Baker has suggested that I supplement the report of the subcommittee on NAEP Writing with some details of my own thinking. This I am happy to do. First, however, I want to say that I am in agreement with the points in the subcommittee report and that my remarks here are intended to provide some of the detailed thinking that supports those points rather than to raise new issues.

My comments are concerned with the following issues: (1) the issue of how objectives of the assessment in writing ought to be stated; (2) the issue of what criteria should guide the selection of exercises for the assessment at each of the several age/grade levels; and (3) the issue of how the results of the assessment ought to be reported. The three issues are closely interrelated, and the separation is only for the purpose of providing a loose framework for my remarks.

(1) Objectives. The Writing Objectives for the 1984 assessment differ from earlier ones in two respects. First, they are stated in general terms without any reference to what might be reasonably expected of individuals in the several age/grade groups or at different levels of ability within age groups. This format provides nothing in the way of specific guides to the choosing of exercises to be used in the assessment. Originally, NAEP was designed to provide for the broad concerned public -- political leaders, parents, educators, and the general public -- pictures of what individuals at various age levels knew or could do. The results were to be presented as directly as possible, by using representative exercises and responses to exercises rather than by summary test scores. There were two reasons for this focus. One was that there was real concern that the typical tests used in the schools, by focusing on the kinds of things that the typical pupil knew or could do, and on the kinds of questions that could be formulated in machine-scorable format, was providing an incomplete assessment of desirable educational outcomes. The other was that the typical test, by summarizing results in test scores, was reporting results that could be properly understood only by people trained in the technology of educational measurement, not by the people responsible for making decisions about educational policy. It seems to me that educational objectives ought to be stated specifically enough to serve as guides to the selection of exercises at the several age/grade levels, not simply in terms of the ultimate objectives of the total educational process. Ideally, the expected outcome statements should be supplemented by actual test exercises and student responses to make clear what the more general statements of objectives are actually implying.

The second way in which the 1984 Writing Objectives differ from earlier ones is that they are stated in terms of process rather than product.

As indicated above, NAEP was originally conceptualized as a program that would communicate to those interested in education what members of selected age groups knew or could do. The focus was clearly on the outcomes of education, i.e. on the products. Given a clear description of what individuals knew or could do, those responsible for making decisions could judge whether or not the results were acceptable and if not could take steps to learn why and take steps to correct the situation. It seems, however, that over the years the actual statements of objectives are being formulated by so-called experts rather than by representatives of the lay public and that NAEP is being designed to provide process data experts can use in exploring their theories rather than to provide the public with descriptions of the outcomes of education in all its variety. I strongly recommend that the question of which type of information is most useful to classroom teachers and lay leaders be raised specifically at the time of the next review of writing objectives. It is my view that NAEP should be concerned with the collection and description of products, leaving the exploration of relationships between educational processes and those products to be done by educational researchers using more complex methodology than the essentially survey methods of NAEP.

(2) Criteria for the Selection of Exercises. The Writing Assessment appears to lack any explicit criteria for the selection of exercises for the several age/grade levels, and many of the actual exercises appear not to be the kinds of exercises that sensitive teachers would expect pupils to undertake in the real world of either the school or in everyday life. For example, in what circumstances, other than as an assignment by some authority figure, would a cross section of nine-year-olds be expected to write a description of a Dali surrealistic painting? The contrast between the exercises included in the 1969-70 assessment of writing and the 1983-84 assessment is dramatic, and it is my judgment that if a representative panel of parents or teachers were to be asked to judge the appropriateness of the exercises, they would favor the 1969-70 set. There appears to be no evidence that the exercises were pretested in any rigorous manner but rather that they were chosen on an impressionistic basis after looking at a limited number of responses collected from whatever samples of individuals were readily available. I recommend that specific criteria be set up in advance for the selection of exercises for the next assessment and that those criteria include the rating of a pretest sample of papers on whatever rating criteria are to be used in the final assessment.

As the draft of the 1984 assessment indicates, "students did produce their responses under restraints of limited time, artificial tests, no access to resourse materials, lack of feedback, and no provision for revising their work at some later time" (Applebee, Langer, & Mullis, July 1986). Thus, one cannot be sure just how valid the responses under these conditions are as indicators of what individuals might do when confronted by real situations requiring written expression. I believe that in designing the next writing assessment program, careful thought should be given to the possibility of collecting data in more realistic contexts, even if it means collecting fewer responses, reducing the opportunities for obtaining data on intercorrelations of exercises, limiting the opportunities for quantifying change across age/grade levels, or using less standardized procedures for collecting responses. Consideration should be given to (a) selecting tasks that will seem more appropriate to selected age/grade groups, (b) offering alternative topics to compensate for the fact that different individuals find different topics legitimate for them, (c) basing

exercises on background material distributed in advance so that individuals might become familiar with it and so have a common background for responding, (d) communicating to the respondents the criteria by which the responses are to be judged, and (e) offering enough time for actual thinking and writing so that there is at least some opportunity for revising and correcting first drafts.

(3) Reporting of Results. On the basis of a review of the report Writing - Trends Across the Decade, 1974-84 and of the draft report The Writing Report Card - Writing Achievement in American Schools, 1984, I conclude that there are serious problems with the reporting of NAEP Writing results and that the design of the next assessment ought to include provisions for changes. More specifically, I believe that: (a) the two rating scales now in use (Primary Trait and Global Rating) should be replaced by anchored rating scales with actual sample responses marking and defining the cutting points between rating levels; (b) more attention should be given to spelling out the relationship of results to stated objectives; (c) more effort should be given to indicating the response patterns for pairs of exercises; (d) the ARM scaling should be abandoned; and (e) attention should be given again to the effect of dependencies across grouping categories on the magnititude of group differences. My reasons for these conclusions follow.

(a) An Anchored Rating Scale. The original rating scale used with NAEP Writing (Global Rating) did a reasonably good job of ordering responses with respect to relative quality, but it did not provide information about how the various scale points were related to desirable outcomes implied by the objectives. The Primary Trait Rating was developed to remedy this limitation. However, studies of the relationships between ratings on the two scales indicate that intercorrelations of ratings are relatively low, that is, the two methods result in a different ordering of papers. One wonders whether it isn't the case that the Primary Trait Rating, by limiting judgments to only certain aspects of the responses, is achieving high reliability at the expense of validity.

There is, however, another way of achieving a quality scale with direct meaning. That is by providing a scale with illustrative responses at the division points between the scale values. It would, of course, be necessary to define directly the qualities that are to characterize the several points on the scale and then to collect and evaluate actual responses at the pretest level -- and to do this for each of the exercises. But such a procedure really ought to be followed anyway in order to insure that the exercises are producing responses of the kinds anticipated. With such a scale, judges would be asked to indicate where each response fell, either between two anchor points or at one of the extremes outside the anchors. The anchors would serve to illustrate for the NAEP public the qualities implied by the numbers on the scales.

It should be made clear to readers of future reports that quality points on such a scale are dependent on the nature of the exercise: what is acceptable performance for a task that might be reasonably required of a fourth grader is not necessarily an acceptable performance for an eleventh grader, and an objective suitable for young children is not therefore also suitable for adolescents.

(b) Relationships to Objectives. One way of indicating relationships of responses on the assessment to the objectives is to spell out more clearly than has been true in the past the relationship of expectations implied by the objectives to the levels of response illustrated by the division points of the scale. But it is also important to spell out the ways in which the exercises are designed to elicit information relative to the several objectives. The reports of the 1984 assessment do this to some extent with respect to Objective II - Students Use Writing to Accomplish a Variety of Purposes. They do not, however, spell out how the responses are related to the other objectives. Either analytical procedures should be developed that tease out such relationships or else specific exercises should be developed to assess these other objectives. Perhaps this task will become more feasible if objectives are more clearly spelled out as suggested in Section 1 of this memorandum.

(c) Response Patterns for Pairs of Items. The original specifications for NAEP Writing called for reporting, exercise by exercise, of responses by very able (top 15%), typical (median), and low ability (bottom 15%) individuals of an age group. The members of the lay public (those not measurement specialists) concerned with educational decisions could then decide whether or not these response patterns were acceptable or whether something should be done to improve performance. Given that individuals responded to only a limited number of exercises, however, it was not possible to show the extent to which those falling in the low group on one exercise would also fall in the low group on a second exercise. Actually, relationships among individual exercises tend to be relatively low, and it seems important that those making judgments about the meaning of an assessment become sensitive to these low relationships, i.e. to understand that just because only 15% of a sample write acceptable responses to one exercise is not sufficient for concluding that only 15% are capable of writing an acceptable response to any exercise. The quality of a response often depends on the extent to which the exercise taps the experience of the individual, and this will differ from exercise to exercise.

When the balanced incomplete block design (BIB) was proposed for the 1984 Writing Assessment, I anticipated that this problem would disappear and that the assessment reports would provide some information regarding these relationships. Instead, I find that the BIB data have been used only to estimate typical performance across a number of different items. I recommend that in future assessments reports include some cross tabulations showing how responses to pairs of exercises are related.

(d) ARM Scaling. Much is made in the 1984 Writing reports of how the BIB design permitted the estimation of average performance across the ten exercises administered to more than one age group. It is my judgment that this attempt to summarize the data is more likely to be misleading than informative. First, the procedure implies that the standard for evaluating responses of young children (age 9) are the same as those for evaluating the responses of older children (ages 13 and 17) or perhaps even of adults. This may very well be the case, assuming that all exercises are equally appropriate for the several age groups. But to set such an examination is to impose a procrustean bed on the evaluation. Better, I think, to limit cross age/grade comparisons to a small number of individual exercises judged explicitly to be appropriate for more than one age/grade group, with no implication that exercises not administered at more than one age/grade might be appropriate for such purposes. Second, it

suggests that the domain of writing can be adequately sampled by ten exercises and that averages across ten such exercises are comparable to those across ten other exercises. My experience with ten-item tests suggests otherwise. Third, it suggests that changes in percentages across age groups are on a common scale with equal units and that differences from 9 to 13 may be compared to differences from 13 to 17.

Now, of course, the measurement specialists who developed the ARM scale know that these kinds of interpretations are not warranted. In fact, they specifically warn against such interpretations. Note, however, that it was to some extent the tendency of lay leaders to misinterpret results of typical standardized tests used in the schools that led to the National Assessment in the first place, and it is unlikely that misinterpretations of ARM scores will be any less likely than misinterpretations of the scaled scores (grade equivalent, normal curve equivalent, etc.) reported for the standardized tests. Actually, the ARM scale used in the 1984 Writing reports rests on much less field research than the typical standardized test scale. If any test publisher had reported test results on a scale with a research base as limited as that for the ARM scale, reviews of competent reviewers would have been scathing. Better, I think, to stick to reports of percentages for individual exercises with the focus on the variety and range of responses, perhaps with averages of the percentages within age/grade samples, perhaps adjusting for extreme effects as was done in earlier NAEP assessments.

(e) Dependencies among Group Effects. I am concerned that the reported group effects may be misleading because of dependencies among the several grouping categories. In earlier NAEP reports, an attempt was made, in reporting group effects, to adjust for such relationships. For example, the effect of race is complicated by the fact that racial samples are not distributed proportionally across types of communities, education of parents, etc. Thus, the actual differences observed in any category are to some extent due to these other factors. The true effects are usually smaller than the observed effects, and this should be made clear in reports.